Close and Distant Reading Visualizations for the Comparative Analysis of Digital Humanities Data

Stefan Jänicke
Dissertation, 2016

Abstract

Traditionally, humanities scholars carrying out research on a specific or on multiple literary work(s) are interested in the analysis of related texts or text passages. But the digital age has opened possibilities for scholars to enhance their traditional workflows. Enabled by digitization projects, humanities scholars can nowadays reach a large number of digitized texts through web portals such as Google Books or Internet Archive. Digital editions exist also for ancient texts; notable examples are PHI Latin Texts and the Perseus Digital Library.

This shift from reading a single book "on paper" to the possibility of browsing many digital texts is one of the origins and principal pillars of the digital humanities domain, which helps developing solutions to handle vast amounts of cultural heritage data – text being the main data type. In contrast to the traditional methods, the digital humanities allow to pose new research questions on cultural heritage datasets. Some of these questions can be answered with existent algorithms and tools provided by the computer science domain, but for other humanities questions scholars need to formulate new methods in collaboration with computer scientists.

Developed in the late 1980s, the digital humanities primarily focused on designing standards to represent cultural heritage data such as the Text Encoding Initiative (TEI) for texts, and to aggregate, digitize and deliver data. In the last years, visualization techniques have gained more and more importance when it comes to analyzing data. For example, Saito introduced her 2010 digital humanities conference paper with: "In recent years, people have tended to be overwhelmed by a vast amount of information in various contexts. Therefore, arguments about 'Information Visualization' as a method to make information easy to comprehend are more than understandable." A major impulse for this trend was given by Franco Moretti. In 2005, he published the book "Graphs, Maps, Trees", in which he proposes so-called distant reading approaches for textual data that steer the traditional way of approaching literature towards a completely new direction. Instead of reading texts in the traditional way – so-called close reading –, he invites to count, to graph and to map them. In other words, to visualize them.

This dissertation presents novel close and distant reading visualization techniques for hitherto unsolved problems. Appropriate visualization techniques have been applied to support basic tasks, e.g., visualizing geospatial metadata to analyze the geographical distribution of cultural heritage data items or using tag clouds to illustrate textual statistics of a historical corpus. In contrast, this dissertation focuses on developing information visualization and visual analytics methods that support investigating research questions that require the comparative analysis of various digital humanities datasets. We first take a look at the state-of-the-art of existing close and distant reading visualizations that have been developed to support humanities scholars working with literary texts. We thereby provide a taxonomy of visualization methods applied to show various aspects of the underlying digital humanities data. We point out open challenges and we present our visualizations designed to support humanities scholars in comparatively analyzing historical datasets. In short, we present (1) GeoTemCo for the comparative visualization of geospatial-temporal data, (2) the two tag cloud designs TagPies and TagSpheres that comparatively visualize faceted textual summaries, (3) TextReuseGrid and TextReuseBrowser to explore re-used text passages among the texts of a corpus, (4) TRAViz for the visualization of textual variation between multiple text editions, and (5) the visual analytics system MusikerProfiling to detect similar musicians to a given musician of interest. Finally, we summarize our and the collaboration experiences of other visualization researchers to emphasize the ingredients required for a successful project in the digital humanities, and we take a look at future challenges in that research field.